Vaccination Rate Mini Project

Soomin Park

Getting Started

Import vaccination data

```
# Import vaccination data
vax <- read.csv("29cd0b19-c7e6-4eb1-8be8-2b6e269f446e.csv")
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction     county
1 2021-01-05                   93704                     Fresno     Fresno
2 2021-01-05                   95684                  El Dorado  El Dorado
3 2021-01-05                   92273                   Imperial   Imperial
4 2021-01-05                   93662                     Fresno     Fresno
5 2021-01-05                   95673                 Sacramento Sacramento
6 2021-01-05                   93668                     Fresno     Fresno
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              2 Healthy Places Index Score
3                              1 Healthy Places Index Score
4                              1 Healthy Places Index Score
5                              2 Healthy Places Index Score
6                              1    CDPH-Derived ZCTA Score
  age12_plus_population age5_plus_population tot_population
1               24803.5                27701          29740
2                2882.9                 3104           3129
3                1633.1                 1763           2010
4               24501.3                28311          30725
5               13671.7                15453          16636
6                1013.4                 1199           1219
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
```

```
2                      NA                            NA
3                      NA                            NA
4                      NA                            NA
5                      NA                            NA
6                      NA                            NA
  percent_of_population_fully_vaccinated
1                                     NA
2                                     NA
3                                     NA
4                                     NA
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                         NA
3                                         NA
4                                         NA
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                     NA                  NA
3                                     NA                  NA
4                                     NA                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        5
2                        NA                        0
3                        NA                        1
4                        NA                        1
5                        NA                        3
6                        NA                        0
  eligible_bivalent_recipient_count
1                                 5
2                                 0
3                                 0
4                                 1
5                                 3
6                                 0
                                                           redacted
1 Information redacted in accordance with CA state privacy requirements
2 Information redacted in accordance with CA state privacy requirements
```

```
3 Information redacted in accordance with CA state privacy requirements
4 Information redacted in accordance with CA state privacy requirements
5 Information redacted in accordance with CA state privacy requirements
6 Information redacted in accordance with CA state privacy requirements
```

```
tail(vax)
```

```
       as_of_date zip_code_tabulation_area local_health_jurisdiction
222259 2023-05-30                   93543               Los Angeles
222260 2023-05-30                   95320               San Joaquin
222261 2023-05-30                   95329                  Tuolumne
222262 2023-05-30                   93517                      Mono
222263 2023-05-30                   95357                 Stanislaus
222264 2023-05-30                   93513                      Inyo
            county vaccine_equity_metric_quartile                 vem_source
222259 Los Angeles                               1 Healthy Places Index Score
222260 San Joaquin                               3 Healthy Places Index Score
222261    Tuolumne                               2 Healthy Places Index Score
222262        Mono                               4    CDPH-Derived ZCTA Score
222263  Stanislaus                               1 Healthy Places Index Score
222264        Inyo                               3 Healthy Places Index Score
       age12_plus_population age5_plus_population tot_population
222259               11902.6                13181          14392
222260               10311.0                11637          12822
222261                2252.1                 2399           2570
222262                 622.3                  639            641
222263                9995.5                11173          11765
222264                1372.5                 1499           1621
       persons_fully_vaccinated persons_partially_vaccinated
222259                     8372                         875
222260                     6977                         559
222261                     1191                         116
222262                      412                          52
222263                     8104                         683
222264                      982                          82
       percent_of_population_fully_vaccinated
222259                               0.581712
222260                               0.544143
222261                               0.463424
222262                               0.642746
222263                               0.688823
```

```
222264                                0.605799
       percent_of_population_partially_vaccinated
222259                                   0.060798
222260                                   0.043597
222261                                   0.045136
222262                                   0.081123
222263                                   0.058054
222264                                   0.050586
       percent_of_population_with_1_plus_dose booster_recip_count
222259                                0.642510                3926
222260                                0.587740                3698
222261                                0.508560                 693
222262                                0.723869                 237
222263                                0.746877                4334
222264                                0.656385                 616
       bivalent_dose_recip_count eligible_recipient_count
222259                      1315                     8369
222260                      1359                     6974
222261                       295                     1190
222262                        94                      411
222263                      1431                     8095
222264                       306                      982
       eligible_bivalent_recipient_count redacted
222259                              8369       No
222260                              6974       No
222261                              1190       No
222262                                 0       No
222263                                 0       No
222264                                 0       No
```

Q1. What column details the total number of people fully vaccinated?

persons_fully_vaccinated

Q2. What column details the Zip code tabulation area?

zip_code_tabulation_area

Q3. What is the earliest date in this dataset?

2021-01-05

Q4. What is the latest date in this dataset?

2023-05-30

```
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	222264
Number of columns	19
_____	
Column type frequency:	
character	5
numeric	14
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
as_of_date	0	1	10	10	0	126	0
local_health_jurisdiction	0	1	0	15	630	62	0
county	0	1	0	15	630	59	0
vem_source	0	1	15	26	0	3	0
redacted	0	1	2	69	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1.00	93665.11	1817.38	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quartile	10962	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.01	18993.87	0	1346.95	13685.13	31756.18	88556.7	
age5_plus_population	0	1.00	20875.24	21105.96	0	1460.50	15364.00	34877.00	101902.0	
tot_population	10836	0.95	23372.72	22628.50	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	17848	0.92	14299.45	15281.94	11	957.00	9034.00	23818.00	67721.0	
persons_partially_vaccinated	17848	0.92	1712.08	2075.03	11	164.00	1204.00	2551.00	43152.0	
percent_of_population_fully_vaccinated	22720	0.90	0.58	0.25	0	0.44	0.62	0.75	1.0	
percent_of_population_partially_vaccinated	22720	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	23383	0.89	0.65	0.24	0	0.50	0.68	0.82	1.0	
booster_recip_count	74543	0.66	6417.22	7795.13	11	331.00	3135.00	10344.00	60058.0	
bivalent_dose_recip_count	160089	0.28	3438.22	4034.61	11	225.00	1863.00	5532.00	29593.0	
eligible_recipient_count	0	1.00	13145.11	15144.22	0	537.00	6691.00	22558.00	87442.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
eligible_bivalent_recipient_0	count	1.00	13038.24	15218.39	0	263.00	6583.00	22550.00	87442.0	

Q5. How many numeric columns are in this dataset?

13

Q6. Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column?

18986

Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)?

.89

Working with dates

```
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
today("2023-06-05")
```

```
Warning in with_tz.default(Sys.time(), tzone): Unrecognized time zone
'2023-06-05'

Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '2023-06-05'

[1] "2023-06-05"
```

```r
# Specify that we are using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

```r
today() - vax$as_of_date[1]
```

```
Time difference of 881 days
```

```r
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
Time difference of 875 days
```

Q9. How many days have passed since the last update of the dataset?

```r
last_update <- ymd("2023-05-30")
current_date <- today()
days_passed <- as.numeric(current_date - last_update)
print(days_passed)
```

```
[1] 6
```

It has been 6 days since the last update of the dataset.

Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)?

```r
# Assuming your dataset is named "data" and the date column is named "date"
unique_dates <- unique(vax$as_of_date)
num_unique_dates <- length(unique_dates)

# Print the result
cat("Number of unique dates in the dataset:", num_unique_dates, "\n")
```

```
Number of unique dates in the dataset: 126
```

There are 126 unique dates in the dataset.

Working with ZIP codes

find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area

```
library(zipcodeR)
```

```
The legacy packages maptools, rgdal, and rgeos, underpinning this package
will retire shortly. Please refer to R-spatial evolution reports on
https://r-spatial.org/r/2023/05/15/evolution4.html for details.
This package is now running under evolution status 0
```

```
geocode_zip('92037')
```

```
# A tibble: 1 x 3
  zipcode   lat   lng
  <chr>   <dbl> <dbl>
1 92037    32.8 -117.
```

Calculate the distance between the centroids of any two ZIP codes in miles, e.g.

```
zip_distance('92037','92109')
```

```
  zipcode_a zipcode_b distance
1     92037     92109     2.33
```

we can pull census data about ZIP code areas (including median household income etc.)

```
reverse_zipcode(c('92037', "92109") )
```

```
# A tibble: 2 x 24
  zipcode zipcode_type major_city post_office_city common_city_list county state
  <chr>   <chr>        <chr>      <chr>                     <blob> <chr>  <chr>
1 92037   Standard     La Jolla   La Jolla, CA           <raw 20 B> San D~ CA
2 92109   Standard     San Diego  San Diego, CA          <raw 21 B> San D~ CA
# i 17 more variables: lat <dbl>, lng <dbl>, timezone <chr>,
#   radius_in_miles <dbl>, area_code_list <blob>, population <int>,
#   population_density <dbl>, land_area_in_sqmi <dbl>,
#   water_area_in_sqmi <dbl>, housing_units <int>,
#   occupied_housing_units <int>, median_home_value <int>,
#   median_household_income <int>, bounds_west <dbl>, bounds_east <dbl>,
#   bounds_north <dbl>, bounds_south <dbl>
```

Focus on the San Diego area

Let's now focus in on the San Diego County area by restricting ourselves first to vax$county == "San Diego" entries

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
sd <- filter(vax, county == "San Diego")

nrow(sd)
```

```
[1] 13482
```

```
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

Q11. How many distinct zip codes are listed for San Diego County?

```
distinct_zip_codes <- vax %>%
  filter(county == "San Diego") %>%
  distinct(zip_code_tabulation_area) %>%
  nrow()
print(distinct_zip_codes)
```

```
[1] 107
```

107 distinct zip codes are listed for San Diego County.

Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset?

```
largest_zip_code <- vax %>%
  filter(county == "San Diego") %>%
  arrange(desc(age12_plus_population)) %>%
  slice(1) %>%
  pull(zip_code_tabulation_area)

print(largest_zip_code)
```

```
[1] 92154
```

Zip Code 92154 has the largest age 12+ population

Using dplyr select all San Diego "county" entries on "as_of_date" "2022-11-15" and use this for the following questions. > Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

```
# Filter the dataset for San Diego "county" entries on "as_of_date" "2022-11-15"
san_diego_entries <- vax %>%
  filter(county == "San Diego" & as_of_date == "2022-11-15")

# Calculate the overall average "Percent of Population Fully Vaccinated"
average_percent_vaccinated <- san_diego_entries %>%
  summarise(average_percent_vaccinated = mean(`percent_of_population_fully_vaccinated`, na

# Print the overall average value
average_percent_vaccinated
```

```
  average_percent_vaccinated
1                  0.7392817
```

Answer: 73.93%

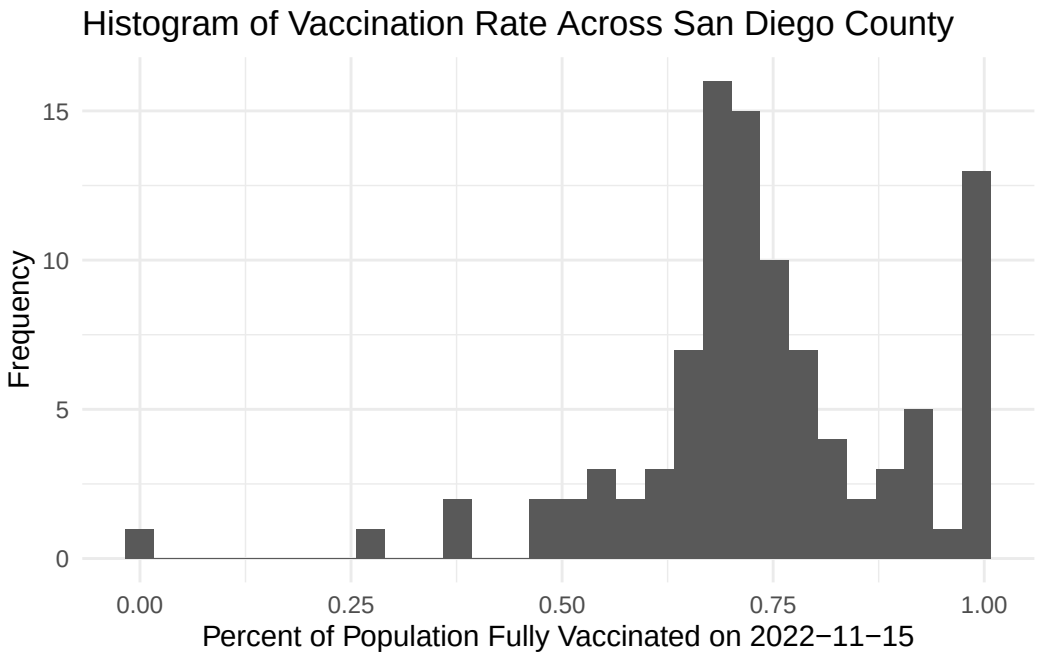
Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2022-11-15"?

```
library(ggplot2)
vaccination_summary <- vax %>%
  filter(county == "San Diego", as_of_date == "2022-11-15")
```

```
ggplot(vaccination_summary, aes(x = `percent_of_population_fully_vaccinated`)) +
  geom_histogram() +
  labs(x = "Percent of Population Fully Vaccinated on 2022-11-15", y = "Frequency") +
  ggtitle("Histogram of Vaccination Rate Across San Diego County") +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 8 rows containing non-finite values (`stat_bin()`).
```



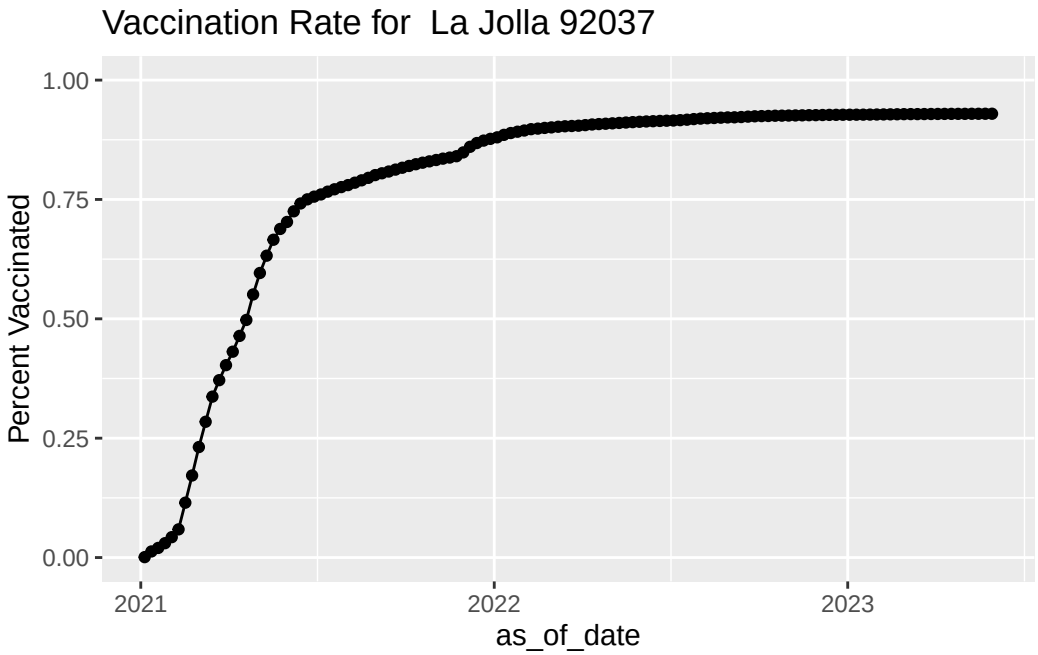
Focus on UCSD/La Jolla

```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

Q15. Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area:

```
ggplot(ucsd) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(title = "Vaccination Rate for  La Jolla 92037", x = "as_of_date", y="Percent Vaccin
```



Comparing to similar sized areas

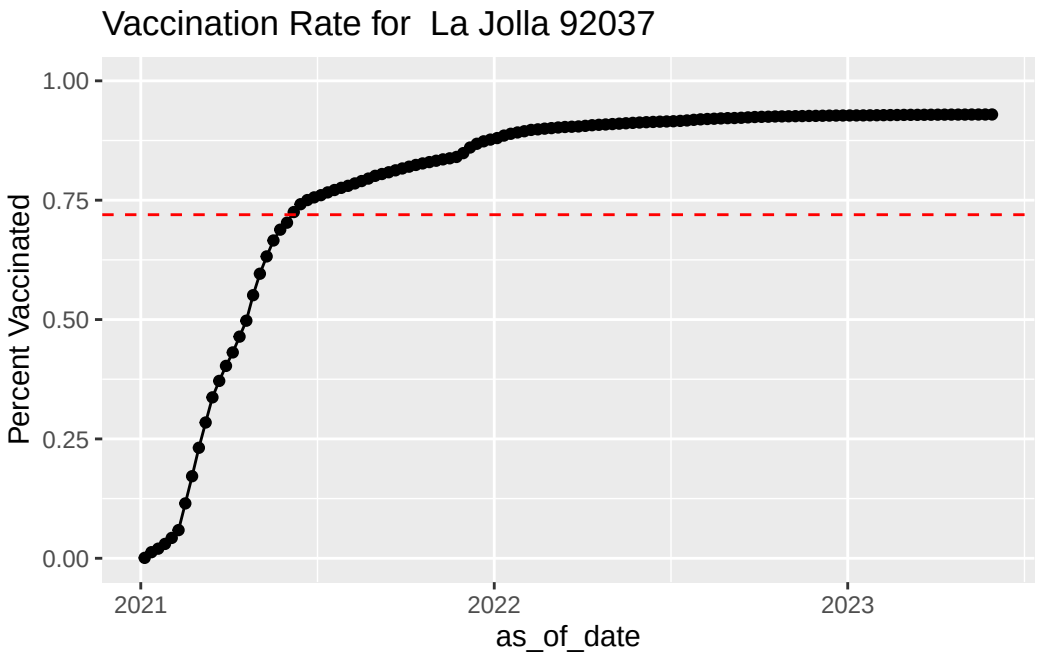
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-11-15")

#head(vax.36)
```

Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```r
# Calculate the mean Percent of Population Fully Vaccinated for ZIP code areas with popula
mean_percent_vaccinated <- vax.36 %>%
  summarise(mean_percent_vaccinated = mean(`percent_of_population_fully_vaccinated`))

ggplot(ucsd) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(title = "Vaccination Rate for  La Jolla 92037", x = "as_of_date", y="Percent Vaccin
  geom_hline(yintercept = mean_percent_vaccinated$mean_percent_vaccinated, linetype = "das
```



Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-11-15"?

```r
# Calculate the 6-number summary
summary_percent_vaccinated <- vax.36 %>%
  summarise(Min = min(`percent_of_population_fully_vaccinated`),
            Q1 = quantile(`percent_of_population_fully_vaccinated`, 0.25),
```

```
            Median = median(`percent_of_population_fully_vaccinated`),
            Mean = mean(`percent_of_population_fully_vaccinated`),
            Q3 = quantile(`percent_of_population_fully_vaccinated`, 0.75),
            Max = max(`percent_of_population_fully_vaccinated`))

# Print the 6-number summary
cat("6-Number Summary of Percent of Population Fully Vaccinated:\n")
```

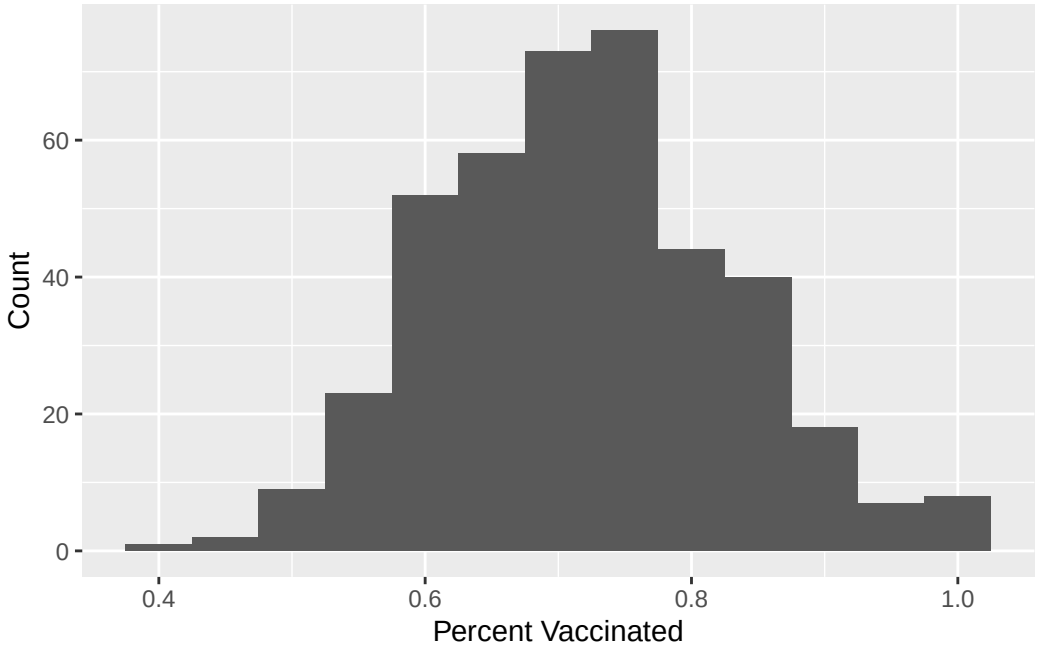
```
6-Number Summary of Percent of Population Fully Vaccinated:
```

```
print(summary_percent_vaccinated)
```

```
      Min       Q1  Median      Mean        Q3 Max
1 0.378957 0.645233 0.71794 0.7196045 0.7896255   1
```

Q18. Using ggplot generate a histogram of this data.

```
# Create the histogram
ggplot(vax.36, aes(x = percent_of_population_fully_vaccinated)) +
  geom_histogram(binwidth = 0.05) +
  labs(x = "Percent Vaccinated",
       y = "Count")
```



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

```
vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.550555
```

```
vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.692471
```

Below the average

Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144.

```r
vax.36.all <- filter(vax, age5_plus_population > 36144)


ggplot(vax.36.all) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2, color="blue") +
  ylim(0, 1) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination rate across California",
       subtitle="Only areas with a population above 36k are shown") +
  geom_hline(yintercept = mean_percent_vaccinated$mean_percent_vaccinated, linetype= "dash
```

```
Warning: Removed 185 rows containing missing values (`geom_line()`).
```